Querying Multiple Bioinformatics Information Sources: Can Semantic Web Research Help?

David Buttler, Matthew Coleman1, Terence Critchlow1, Renato Fileto, Wei Han, Ling Liu, Calton Pu, Daniel Rocco, Li Xiong

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

December 2002

DISCLAIMER

Querying Multiple Bioinformatics Information Sources: Can Semantic Web Research Help?

David Buttler, Matthew Coleman[1], Terence Critchlow[1], Renato Fileto, Wei Han,
Ling Liu, Calton Pu, Daniel Rocco, Li Xiong

Abstract *Advances in Semantic Web and Ontologies have pushed the role of semantics to a new frontier: Semantic Composition of Web Services. A good example of such compositions is the querying of multiple bioinformatics data sources. Supporting effective querying over a large collection of bioinformatics data sources presents a number of unique challenges. First, queries over bioinformatics data sources are often complex associative queries over multiple Web documents. Most associations are defined by string matching of textual fragments in two documents. Second, most of the queries required by Genomics researchers involve complex data extraction, and sophisticated workflows that implement the complex associative access. Third but not the least, complex Genomics-specific queries are often reused many times by other Genomics researchers, either directly or through some refinements, and are considered as a part of the research results by Genomics researchers. In this short article we present a list of challenging issues in supporting effective querying over bioinformatics data sources and illustrate them through a selection of representative search scenarios provided by biologists. We end the article with a discussion on how the state-of-art research and technological development in Semantic Web, Ontology, Internet Data Management, and Internet Computing Systems can help addressing these issues.*

1 Introduction

Many biologists are highly motivated to make the data generated from their research available on the web, either because of funding requirements, recognition, or to push the frontiers of their science. However, while a lot of data resides in several large repositories, it is not accessible in a machine-processable format, which restricts how it can be used by other researchers. First, there are hundreds of tools [6] available online to analyze data, each with its own interface supporting slightly different invocation, processing and data semantics. This makes using the tools difficult because the input

must be guaranteed to be in the correct format with the correct semantics and the tools must be invoked in specific, non-standard ways. Second, cutting edge work can be made available directly from the research lab where it is created long before it ends up in the large repositories; in fast moving research environments having access to the latest work is essential. Third, even the data at well-known sources, such as NCBI or EMBL, is not necessarily available for in-depth analysis primarily because the interfaces provided involve human interaction. Data gathering and analysis tasks then require a significant amount of time to enter requests and retrieve the results. For high-throughput experiments this can become a significant bottleneck. The Semantic Web provides an opportunity to remove this bottleneck and enable seamless interoperation of resources. Automating query and analysis tools can revolutionize the way that biologists currently use the information in research and development.

2 Motivation

One effective way to understand the issues facing bioinformatics is to analyze the problems faced by genomics researchers on a daily basis. A careful analysis in this scenario-based approach reveals many opportunities where further research in computer science problems can pay a large dividend in the quality of genomics and other biological research, as well as the quantity of results and the speed at which new research can be proposed, understood, and accomplished. Here we present some background and several scenarios to motivate and explicate research issues related to the pursuit of bioinformatics. For purposes of explanation, the biology has been significantly simplified. Please consult the references for more detailed descriptions.

Scenario Description

Biologists are currently using a variety of tools, such as DNA microarrays, to discover how DNA and the proteins, they encode allow an organism to respond to various stress conditions such as exposure to

environmental mutagens [15, 1, 6]. To do this researchers identify genes that react in the desired way, then develop models to capture the common elements. This model can then be used to identify previously unidentified genes that may also respond in similar fashion based on the common elements.

Figure 1 represents the workflow that a genomics research has created to gather the data required for this analysis. This type of workflow significantly differs from traditional workflows, as it was iteratively generated to discover the correct process using a small set of data – after each step the researcher selected what part of the output data is useful for the next step, and which services the data should be sent to next. Once the workflow was constructed, it is used to analyze large quantities of data. Because each step may require a significant amount of time, the entire process must proceed without human attention after it has been constructed.

In the first step of the workflow, microarrays containing the genes of interest are produced and exposed to different levels of a specific mutagen in the wet-lab, usually in a time dependent manner.
Gene expression changes are then measured and clustered with computational tools. The researcher must choose from a wide variety of tools available for this task. Each tool offers specific advantages in terms of their ability to analyze the microarray data, and each requires a different method of execution. As more tools come online that offer different methods of analysis, the problem of choosing which tools to use for this step, running the data through the tool, and gathering the analysis requires semantic techniques to automate tool (service) selection and activation.

The third step is to retrieve the full sequence from the gene id's chosen in the second step. This information is widely replicated at many different sites, allowing biologists to choose which source to retrieve the information from. However, few people are aware of the mirrors, so well known resources, such as GenBank [8], are typically overloaded with requests during peak hours. In any automated solution the ability to discover new services – either mirrors of a known service or other services in the same domain – and select an equivalent source that has a lower cost of retrieval not only speeds up the

local request, but provides load balancing across the known set of mirrors.

Fourth, once the complete sequence for a relevant gene has been retrieved, it is sent to a gene matching service that will return homologs, other genes with similar sequences. Again, several sites provide gene similarity matching, many of which specialize in a particular species, such as ACEdb [18]. Choosing the appropriate source depends on the content, capabilities and load of the source, as well as the trustworthiness of the source. Some sites have much stricter standards on the quality of the data that they admit, while others publish information as soon as it is available. Depending on the current needs of a particular researcher, different types of sites may be more appropriate to query.

In addition to selecting a capable and trustworthy source, there are significant issues in extracting data from the sites. Most sites have custom query interfaces and return results through a series of HTML pages. For example, NCBI BLAST (Basic Local Alignment Search Tool) [1] requires three or four steps to retrieve sequence homologs. First, a gene sequence must be submitted through an HTML form. Users may then optionally select the format that the data should be returned as. Then, a series of delay pages are shown while the service calculates the final answer. Once the answer is computed, a page listing the related sequence ids and their alignment information is presented. The full homolog sequence is available by following a link from each alignment. Just to retrieve one set of similar sequences from this tool requires a significant amount of human effort in following each link and merging data from the final result pages.

Once related sequences are discovered, approximately 1000-5000 bases of the DNA sequence around the alignment is extracted to capture the promoter regulatory elements -- the region of a gene where RNA polymerase can bind and begin transcription to create the proteins that regulate cell function.
In the fifth step, these promoter sequences are identified and analyzed using specific tools, such as Mat-Inspector [14], TRANSFAC, TRRD, or COMPEL [15] to find the common transcription binding factors. To extract specific data, such as portions of a DNA sequence, returned by sources,

the data needs to be converted into a well-known format, such as XML, and post-processed to extract just the portions that are relevant for the next step.

Once found in step six, regulatory profiles are then compared across each gene in the cluster to delineate common response elements that can be used as a promoter model, developed in step seven. Once the model is created, it can be used to search gene databases to find other candidate genes relevant to the study. These genes can be fed back into the general workflow to refine and expand the promoter model or presented as the final results for the analysis task. Each of these steps requires the same level of multi-source integration, automated data extraction, and semantic integration as the previous steps.
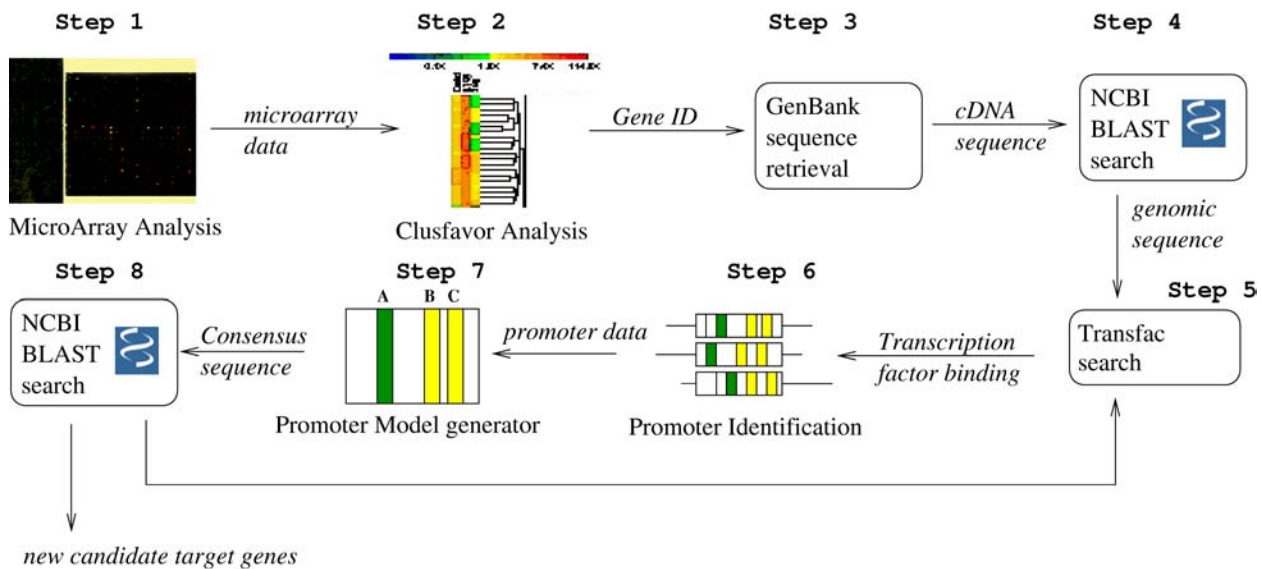

Figure 1 -- Developing a Promoter Model

3 Research Challenges

The Semantic Web provides an excellent opportunity to continue the distributed development of resources while providing a mechanism where independent components can be combined into complex applications. However, at this point the bioinformatics community has not embraced the capabilities provided by the Semantic Web. To understand how the community can benefit with these capabilities, we examine several current research areas and how the Semantic Web impacts their effectiveness. We examine issues in workflows, resource discovery and selection, data provenance and trust, as well as data extraction and integration. Our survey is not exhaustive and there are many important issues and challenges not listed here.

3.1 Process Management

As our scenario demonstrates, scientists need to perform a set of tasks having complex dependencies, requiring that they be done in a particular sequence. In order to manage these processes, several researchers are investigating the use of workflows to increase the quality, reliability, accuracy, understandability and productivity of research efforts [11, 13, 19].

Workflows organize interoperating and potentially distributed data processing activities to facilitate cooperation in achieving some goal [4, 19]. One of the major differences between scientific workflows and traditional workflows is that the process is discovery driven as opposed to codifying rules by which an organization is run. This leads to different

expectations of the workflows flexibility, adaptability, and construction. The Semantic Web enables workflow technology by providing standard mechanisms for description of service capabilities and content.

3.2 Resource Discovery

The value of accessing data from other institutions and the relative ease of disseminating this data has caused an increase in both the capacity for collaboration and the amount of available bioinformatics information dramatically.

Source Discovery

Source discovery is the process of automatically locating a data source and discovering its capabilities and the type of data that it contains. The challenge is to be able to determine when a particular source matches a more generic type of source. For example, the NCBI BLAST interface is a specific instantiation of a generic BLAST search. While it shares much in common with other BLAST searches, the interface is not identical.

One approach to the problems of discovery and classification of bioinformatics sources is the notion of a service class description. A service class description encapsulates the relevant portions of the service from the point of view of the intended application. The description includes the various data types used by the service, example queries and output, and a graph representation of how service class members are expected to operate. For example, a simplified view of the DNA sequence BLAST service class includes a DNA sequence input type, a DNA BLAST result output type, and descriptions of the intermediate pages. The control flow graph might show the input page being connected to a result page, possibly through a delay page.

The service class description provides an abstract view of bioinformatics sources that allow developers to reason about the class as a whole, freeing them from concerns about the intricate details of each member. However, it magnifies issues of trust and reliability, discussed below. As data sources and tools move towards publishing their interfaces in a standard format, using standard ontologies, source categorization will become more accurate and automatic.

Source Selection

In our scenario, the biologist must know in advance which tool to use for each step in this process. One problem with this approach is that new tools have a hard time gaining acceptance because they are difficult to discover, and it is difficult to integrate them into an existing process. Another problem is that popular sites become overloaded with requests, even if there are several mirrors willing to share the load, because the mirrors are not well known.

For each stage in gathering the information and building the promoter models, automated source selection gives the biologist more choices and an opportunity to gather richer information more quickly. The Semantic Web is now providing means for uniform publication of resources, allowing automated decision processes to select appropriate sources for each query, based on the requirements of individual researchers.

3.3 Quality of Data

Medical research requires tight controls on the quality of data because mistakes can harm people's health. Data quality in bioinformatics may not be as immediate, but it is no less important. As scientists pursue their research agendas they must remain vigilant on the quality of their data, otherwise their results could be invalidated. When reusing others results, the term quality includes issues of trust in particular data sources, methods, and research institutions.

Trust

In the context of bioinformatics, trustworthiness should capture the consistency, reliability, competence, and honesty of the data source. With the increasing number, complexity and sometimes uncertainty of available data sources, a major barrier to the efficient access to these data sources is their trustworthiness. It is very important to choose reliable and reputable data sources and filter out the unreliable ones.

The Semantic Web provides the means for machine-processable descriptions of data, communications of the trust information, and thus automates much of the filtering process. Each data source could be associated with a trust value, which, similar to the E-Commerce domain [17], can be a result of one or a

combination of self-descriptions of the data sources, community ratings collected by independent rating services, and certifications from trusted third-party authorities. The trust value could be stored in independent trusted bureaus and in turn accessed by the context-understanding programs such as data integration engines, data source discovery programs to select trustworthy data sources. The challenges remain as to how to associate effective trust values with the data sources and how to counter the potential vulnerabilities and threats.

Data Provenance

Data provenance is the recorded history of where a piece of data has come from, and all of the transformations, corrections, and annotations that have occurred to it since creation. In bioinformatics research, data has been collected from many different institutions, using multiple methods; some of the largest databases include data that have been curated multiple times. These characteristics make data provenance extremely important. One technique researchers are pursuing is to introduce a system that annotates data to indicate its provenance [3]. These elements include why, the source data that influenced the existence of the annotated element, as well as where, the locations from which the annotated element was extracted.

3.4 Data Integration

Data integration introduces several complex information management challenges. First, data must be extracted from multiple sources, each of which uses an evolving custom data format. Second, the extracted data must be normalized into a consistent syntactic representation and semantic frame of reference. Only after both of these steps have been accomplished can services (both data sources and analysis programs) be connected in complex workflows to solve a specific problem or answer detailed requests.

Data Extraction

As the Semantic Web develops, more and more information will be presented as Web Services, described by a formal language such as WSDL [4], however, the vast majority of bioinformatics information that is available online, and is coming online in the near future, is in HTML. Wrappers have been key tools to make the conversion from HTML into semantically meaningful and well-structured

XML data. However, developing wrappers has been slow and tedious work with typically brittle results. Developing robust wrappers requires overcoming at least three challenging issues. First, wrappers often require domain specific knowledge from domain experts or end-users. It is important to develop a methodology that separate domain specific knowledge from the wrapper implementation. This separation allows wrappers to be generated for a particular data source given a complete description of the data semantics and interface. Second, a wrapper should be able to tolerate some slight changes in Web pages or be able to be updated easily. Separating semantics from implementation shields wrappers from minor changes, as any change that causes a wrapper to break can trigger automatic regeneration. Third, a wrapper generation system should produce as efficient wrappers as hand-coded ones.

Semantic Integration – Metadata and Ontologies

Once services have been connected through a workflow, the data output from each step must be transformed so that its syntax and semantics match the input expectations for succeeding steps. Metadata can describe both the capabilities and contents of a particular data source [9]. Ontologies allow the evaluation of the meaning of the terms used in data and metadata, by organizing these terms according to their semantic relationships (e.g. holonym, hypernym, etc.). This mechanism is the basis for identifying the relevant data elements of individual sources for particular applications, and the semantic relationships among these data elements, in order to map heterogeneous data fragments into a common frame of reference, enabling the correct mix of data from different sources. Most bioinformatics data sources require significant amounts of effort on the semantic aspects of data integration. There are many different data formats, using different layouts and terminology. While there are some links between the most popular sources (e.g. links from NCBI BLAST results to a related PDB [2] protein structure), these links are not consistent across sources.

The largest hope for fulfilling the ambitions of the Semantic Web in the bioinformatics realm is the definition of a concise and widely accepted ontology for the community. Unfortunately, there is little consensus on many terms, so this solution will be years away. In the meantime, partial ontologies

defined by individual institutions can be extremely useful. Integrating these individual ontologies is known to be hard [10, 12], but their existence allows some meaningful information to be automatically shared, without requiring a human expert to interpret every piece of information.

4 Conclusion

The goal of this study is to identify challenges in the discovery, search, and access of multiple bioinformatics data sources that can be addressed by using the Semantic Web in conjunction with database technologies and techniques. Rather than presenting new problems, our approach has been to examine issues that real biologists are facing and identify current computing research that can drive solutions to the problems that we have perceived. While there is great potential for the Semantic Web to alleviate many of the problems discussed here, there still needs to be significant effort on the part of those institutions which host the data and applications to provide the information necessary to enable the appropriate semantic tools. Until that time, there is still an opportunity for third party efforts to provide the tools and semantic information that can integrate the immense amount of critical information that is already available.

References

1. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research* **25** (1997) 3389-3402

2. H. Berman *et al.*, Protein Structures: From Famine to Feast *American Scientist* **90.4** (2002) 350 - 359.

3. P. Buneman *et al.,* Why and Where: A Characterization of Data Provenance *International Conference on Database Theory (ICDT)* (2001)

4. Z. Cheng *et al.*, Composition Constraints for Semantic Web Services, *WWW2002 Workshop on Real World RDF and Semantic Web Applications* (2002)

5. E. Christensen, *et al.*, Web Services Description Language (WSDL) 1.1, Technical Report, World Wide Web Consortium (2001). See http://www.w3.org/TR/wsdl

6. DBCAT, The Public Catalog of Databases. See http://www.infobiogen.fr/services/dbcat/

7. A.J. Fornace *et al.*, The complexity of radiation stress responses: analysis by informatics and functional genomics approaches, *Gene Expr* **7** (1999) 387-400.

8. GenBank, *Nucleic Acids Research* **30(1)** (2002) 17-20

9. C.A. Goble *et al.*, Transparent access to multiple bioinformatics information sources, *IBM Systems Journal* **40.2** (2001) 532-551

10. A. Gupta *et al.*, Registering Scientific Information Sources for Semantic Mediation, *21st International Conference on Conceptual Modeling,* (2002).

11. K. J. Kochut *et al.*, IntelliGEN: a distributed workflow system for discovering protein-protein interactions, *International Journal on Distributed and Parallel Databases, Special Issue on Bioinformatics* (2002)

12. B. Ludäscher *et al.*, Model-Based Mediation with Domain Maps, *17th Intl. Conference on Data Engineering* (2001)

13. J. Meidanis *et al.*, Using Workflow Management in DNA Sequencing, *Intl. Conf. on Cooperative Information Systems* (1996), 114-123

14. L. Peterson, CLUSFAVOR, *Baylor College of Medicine* (2002). See http://mbcr.bcm.tmc.edu/genepi/

15. K. Quandt *et al.*, MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucl. Acids Res.* **23** (1995) 4878-4884.

16. Z. Ronai, Deciphering the mammalian stress response - a stressful task, *Oncogene* **18** (1999) 6084-6.

17. M. Shepherd *et al*., Building Trust for E-Commerce: Collaborating Label Bureaus, *ISEC 2001, LNCS 2040* (2001) 42-56

18. L. Stein *et al.,* Scriptable access to the Caenorhabditis elegans genome sequence and other **ACEDB** databases. *Genome Res.* **8** (1999) 1308-1315

19. G. Wiederhold *et al.*, Composing Diverse Ontologies, *Technical Report, Stanford University* (1998)

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551